Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Kierunek: **Informatyka Techniczna (ITE)**
Specjalność: **Inżynieria Systemów Informatycznych (INS)**

PRACA DYPLOMOWA MAGISTERSKA

Wykorzystanie algorytmów genetycznych w systemach wykrywania intruzów w sieciach komputerowych

inż. Bartosz Błyszcz

Opiekun pracy
dr inż. Tomasz Babczyński

Słowa kluczowe: 3-6 słów

WROCŁAW, 2024

Streszczenie

Wykaz skrótów

Tabela 1. Tabela skrótów
Źródło: opracowanie własne

GA	*Genetic Algorithm*	Algorytm Genetyczny
GP	*Genetic Programming*	Programowanie Genetyczne
GNB	*Gaussian Naive Bayes*	Naiwny Klasyfikator Bayesa wykorzystujący rozkład Gaussa
ANN	*Artificial Neural Network*	Sztuczna sieć neuronowa
CNN	*Convolutional Neural Network*	Konwolucyjna sieć neuronowa
ML	*Machine Learning*	Uczenie maszynowe
AI	*Artificial Intelligence*	Sztuczna Inteligencja
IDS	*Intrusion Detection System*	System Wykrywania Intruzów
SVM	*Support Vector Machine*	Maszyna Wektorów Nośnych

Spis treści

1. Wstęp

1.1. Wprowadzenie i uzasadnienie tematu pracy

Klasyfikacja danych tabelarycznych jest zagadnieniem, które na codzień dostarcza wyzwań jej twórcom z powodu mnogości danych, a także mnogości cech, a także z nierzadko małą ilością próbek. Jednym z problemów jest między innymi dobór odpowiedniego algorytmu do problemu. Dane tebalryczne występują w każdej dziedzinie, przez co raz na jakiś czas proponowane są nowe rozwiązania i algorytmy mające rozwiązać problem klasyfikacji w sposób lepszy i wydajny. Częśc twórców próbuje podchodzić do tego w sposób innowacyjny, lecz nie zawsze to wychodzi z powodu chociażby dosotoswania algorytmu pod konkretną strukturę danych, co powoduje problemy z wykorzystaniem rozwiązania dla innych danych.

Obecnie jednymi z najpopularniejszych algorytmów do klasyfikacji danych są logiczna regresja(*ang. logistic regression*), drzewo decyzyjne(*ang. decision tree*), losowy las(*ang. random forest*), maszyna wektorów nośnych(*ang. support vector machine*), naiwny bayes(*ang. Naive Bayes*). Dlatego też bardzo ważne jest porównanie wytworzonego wcześniej rozwiązania z grupą innych algorymów, które próbują przetworzyć ten sam zestaw danych.

W dzisiejszych czasach próba taka jest bardzo uproszczona chociażby przez takie platformy jak *Machine Learning Studio*, które pozwalają na wykorzystanie mocy obliczeniowej sklasteryzowanych jednostek wirtualnych do wykonywania obliczeń na odpowiednich maszynach wirtualnych, a także do budowania skomplikowanych zautomatyzowanych procesów złożonych z wielu zadań(*ang. pipeline*). W związku z czym możliwośc wykorzystania platformy chmurowej pozwoli na zautomatyzowanie procesu porównawczego oraz oddelegowanie zadań od chmury obliczeniowej co pozwoli na uniezależnienie powodzenia doświadczenia od mocy obliczeniowej komputera lokalnego, a także na ukazanie całościowo procesu porównania algorytmów klasyfikacyjnych.

1.2. Cel pracy dyplomowej

Celem niniejszej pracy dyplomowej jest porównanie algorytmu klasyfikacji danych tabelarycznych wypracowanego w trakcie pisania pracy inżynierskiej, do algorytmów dostępnych w aplikacji *Machine Learning Studio* znajdującej się na platformie *Microsoft Azure*.

1.3. Założenie techniczne

Dane prezentowane w **tabeli 1.1** określają podstawowe założenia techniczne przyjęte w trakcie wykonywania analizy porównawczej. Dane te dotyczą między innymi środowiska, w którym wykonane było doświadczenie. Dodatkowo uwzględniono zestaw danych oraz biblioteki użyte w trakcie tworzenia doświadczenia.

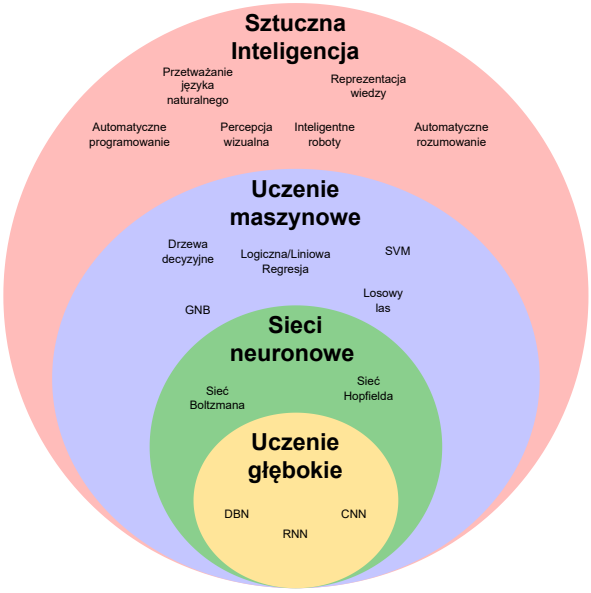
Tabela 1.1. Założenia techniczne pracy dyplomowej
Źródło: Opracowanie własne

Środowisko uruchomieniowe	Machine Learning Studio[1]
Język oporogramowania	Python 3.x
Wykorzystane biblioteki	scikit-learn [**sckit-learn**]
	Numpy [2]
	Pandas [3, 4]
Wykorzystane dane	CICDS2017 [5]

2. Sztuczna inteligencja

Według słownika *Oxford English Dictionary* słowo ”**inteligencja**” oznacza zdolność do rozumienia, a analizy i dostosowania się do zmian[6].

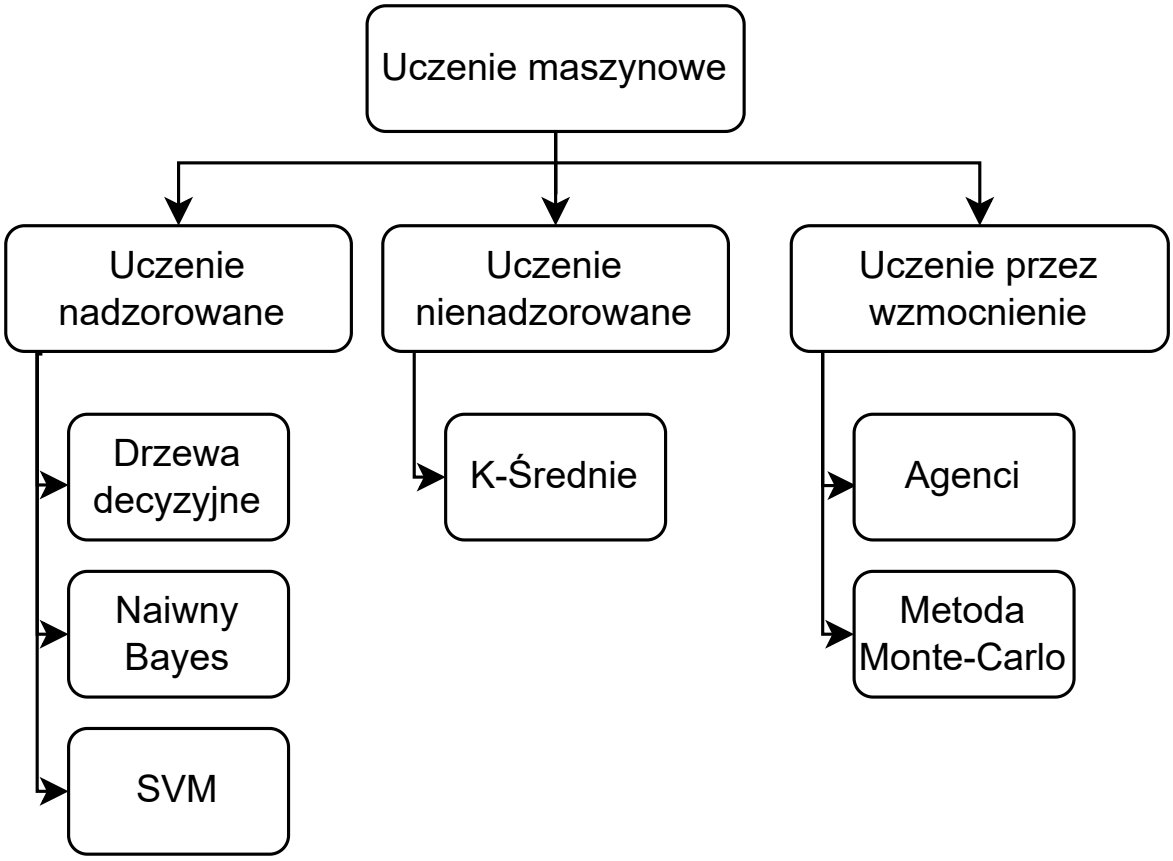
Sztuczna inteligencja(*ang. Artificial Intelligence*) (**AI**) jest wykorzystywana na wiele sposobów podczas prowadzenia badań naukowych: od stawiania hipotez oraz budowania twierdzeń matematycznych, tworzenia i monitorowania badań, zbierania danych i wielu innych czynności towarzyszącymi podczas badań. Najpopularniejszymi zastosowaniami jest między innymi uczenie nienadzorowane oraz wykrywanie anomalii[7, 8]. Schemat podziału sztucznej inteligencji pokazano na **obrazie 2.1** .



Rys. 2.1. Graficzne przedstawienie podziałów sztucznej inteligencji
Źródło: [9]

2.1. Uczenie maszynowe

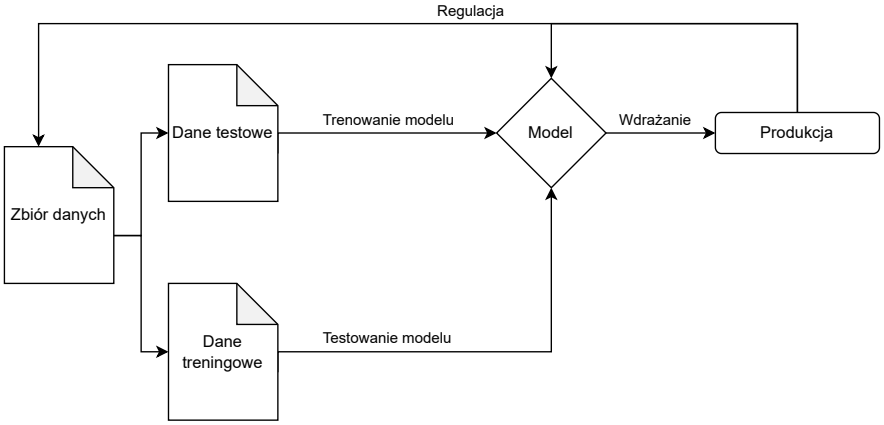
Uczenie maszynowe(*ang. Machine Learning*) (**ML**) jest to dziedzina nauki nad algorytmami oraz modelami statystycznymi, które mogą być wykorzystywane do specyficznych zadań na przykład klasyfikacji, rozpoznawania obrazów bądź mowy, a dodatkowo nie są zaprogramowane specyficznie pod konkretne zadanie, a jedynie pod grupę zadań tak jak pokazano na **obrazie 2.2** . Dlatego też nie ma jednego najlepszego rozwiązania, które można wykorzystać w każdym przypadku. Wykorzystanie konkretnego algorytmu determinuje typ zadania jaki ma być rozwiązany.



Rys. 2.2. Podział uczenia maszynowego
Źródło: [8]

2.1.1. Uczenie nadzorowane

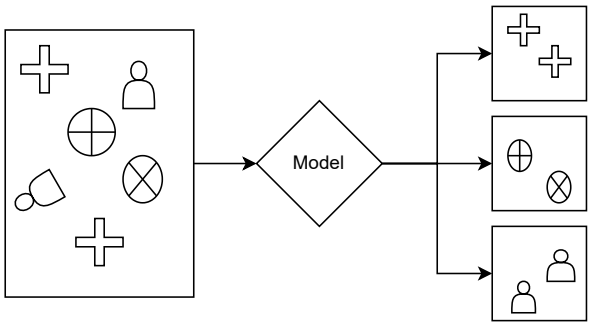
- w trakcie tego uczenia stosuje się zbiór posiadający etykiety. Model uczy się przyporządkowywać określone cechy do konkretnych kategori. Dane wejściowe dzielone są na dane treningowe i dane testowe. Zbiór treningowy jest wykorzystywany do trenowania modelu, a zbió testowy do sprawdzenia rezultatu, na bazie którego może nastąpić korekta uczeniam zilustrowano to **obrazem 2.3** . Algorytmy uczenia nadzorowanego można zastosować między innymi do weryfikacji ruchu sieciowego w celu określenia czy ruch bezpieczny, przez co można to zastosować w systemach wykrywania intruzów(*ang. Intrusion Detection System*) (**IDS**). Algorytmy wchodzące w skład uczenia nadzorowanego to między innymi klasyfikacja naiwna bayesa, drzewa decyzyjne, maszyny wektorów nośnych[7, 8].



Rys. 2.3. Uczenie nadzorowane
Źródło: [8]

2.1.2. Uczenie nienadzorowane

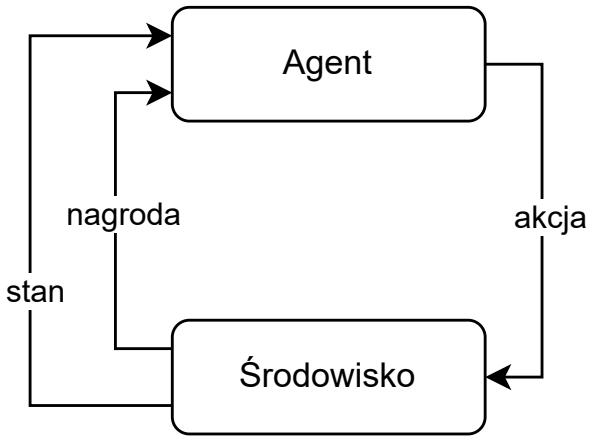
- w tym przypadku nie wykorzystuje się zbioru oznaczonego, algorytm sam próbuje odkryć prawidłową odpowiedź. Dzieję się tak, w danych, których nie da się nazwać albo doprecyzować. Wykorzystuje się to do między innymi detekcji anomalii, co pozwoli do na przykład wykrycia zbyt dużego zużycia prądu w pokoju domu studenckiego dzieki czemu uda się wyłapać nieautoryzowaną koparkę kryptowalut. Dodatkowo można wykorzystać je do szukanai wzorców, albo zarządzania magazynem. W skład takich algorytmów wchodzi: K-średnie, klasteryzacja. Schemat uczenia nienadzorowanego poprzez klasteryzację jest pokazany na **obrazie 2.4** .



Rys. 2.4. Uczenie nienadzorowane
Źródło: Opracowanie własne

2.1.3. Uczenie przez wzmocnienie

- jest to uczenie poprzez nagradzanie dobrych rozwiązań, a karanie złych, potocznie mówiąc jest to metoda ”*kija i marchewki*”. Wykorzystywana w trenowaniu pojazdów autonomicznych pozwala na nagradzanie pojazdów za wybór lepszych tras przykładowo za wybór dróg asfaltowych zamiast polnych. Skupia się w dużym stopniu na agencie i jego decyzjach w danym środowisku co pokazano na **schemacie 2.5** . Należy do jednych z trzech głównych paradygmatów obok uczenia nadzorowanego i nienadzorowanego.



Rys. 2.5. Uczenie przez wzmocnienie
Źródło: [8]

2.1.4. Uczenie częściowo nadzorowane

- do trenowania takich modeli stosuje się niewielkie zbiory oznaczone, oraz większe zbiory nieoznaczone, dzięki którym można próbować rozpoznać rozległe zbiory danych na podstawie pewnych cech wspólnych. Stosuje się to ze względu na mnogość danych na świecie, których opisanie byłoby niemożliwe oraz albo zbyt kosztowne. Przykładowo można znaleźć zastosowanie tych algorytmó w bankowości albo klasyfikowaniu stron internetowych poprzez wyszukiwanie treści na stronie i kategoryzowaniu ich[10]. Jest to połączenie uczenia nadzorowanego i nienadzorowanego[8].

2.2. Sieć neuronowa

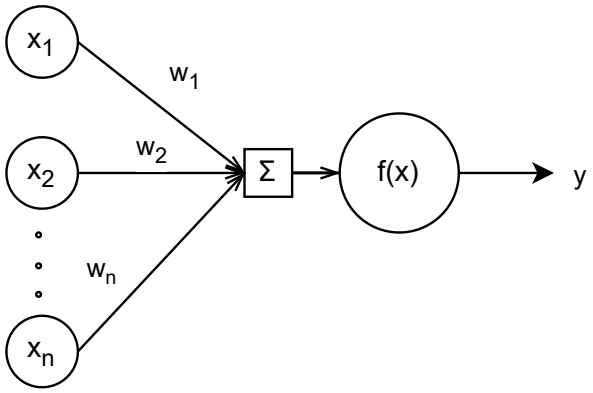
Ludzki mózg jest najbardziej złożonym organem znany ludziom. Badacze zainspirowani jego strukturą składającą się z połączonych ze sobą komórek neuronowych, które przetwarzają równolegle wiele informacji, próbują przenieść pewien poziom inteligencji do komputerów. Przykładem tego jest wiele algorytmów, wchodzących w skład sztucznych sieci neuronowych(*ang. artificial neural network*) (**ANN**), między innymi sieci Kohonena, sieci Hopfielda, sieci konwolucyjne. Sieci te próbują w pewien sposób odwzorować próbę na przykład klasyfikacji danych przez jednostkę wzorowaną na ludzkim mózgu, pomimo tych osiągnięć symulacja ludzkiej świadomości oraz emocji wciąż jest jedynie w sferach fantazji naukowych[11].

Sieć neuronowa jest zbudowana z połaczonych ze sobą warstw neuronów tak jak na **rysunku 2.7** , które w pewien sposób mają wykonać zadania uczenia maszynowego. Najprostszym przykładem sieci neuronowej jest pojedynczy neuron, który może służyć do prostych zadań klasyfikacyjnych: **schemat 2.6** . Sieć ta potrafi się dostosowywać do danych wejściowych tak aby uzyskać odpowiedni wynik, wykonuje w tedy proces uczenia stosując do tego na przykład algorytm wstecznej propagacji wag. W zależności od problemu istnieje wiele różnych sieci, które można zastosować. Jednym z trudniejszych rzeczy w doborze sieci jest dobór warstw ukrytych oraz ilości neuronów, ponieważ w tym celu twórca może opierać się jedynie na własnej wiedzy i doświadczeniu. Podstawy teorii sieci neuronowych zostały stworzone w połowie XX wieku. Złota era uczenia maszynowego rozpoczęła

się dopiero na początku XXI wieku, kiedy to jednocześnie pojawiły się takie trendy jak: Big Data, redukcja kosztów obliczeń równoległych, oraz pierwsze badania nad głębokimi sieciami neuronowymi(*ang. Deep Neural Network*) (**DNN**). Największe zastosowanie DNN miało miejsce dopiero w ostatniej dekadzie kiedy to pojawiły się:

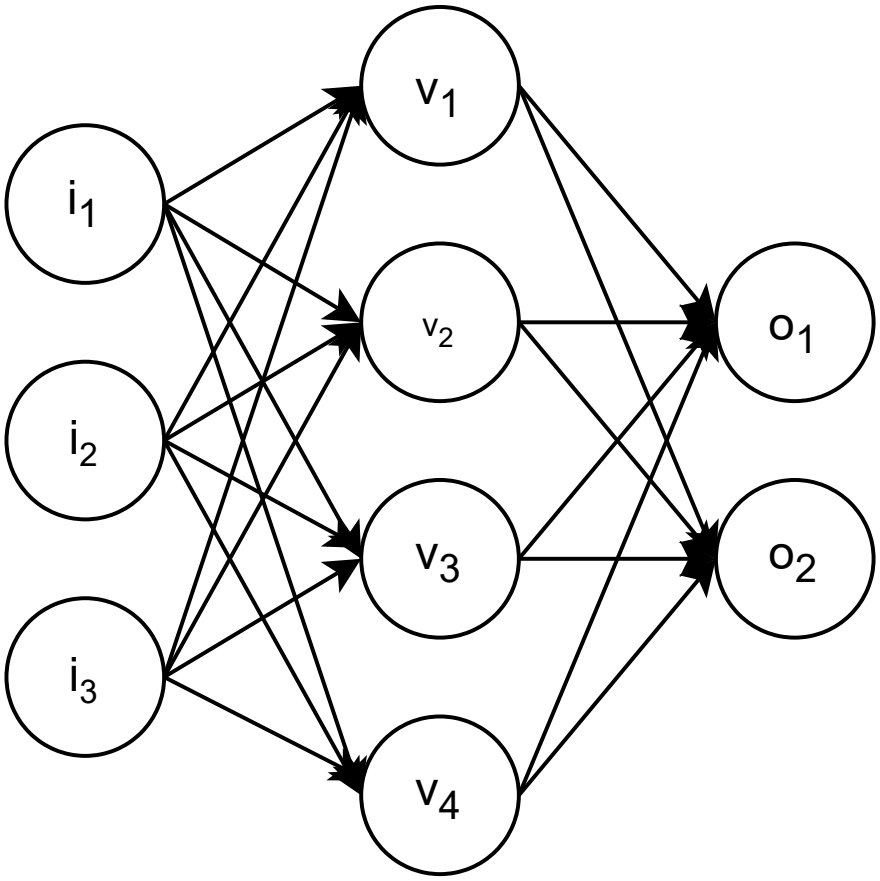
– **Google Braine** - grupa badawcza założona w 2011 roku, zajmująca się badaniami nad sztuczną inteligencja

– **DeepFace** - rozwiązanie stworzone przez firmę Facebook w 2014 roku, służące do rozpoznawania twarzy na zdjęciu [12, 13].



Σ: sumator
$f(x)$: funkcja aktywacyjna
$w_x \; \forall x \; \in [1, 2, ..., n]$: wagi
$x_x \; \forall x \; \in [1, 2, ..., n]$: wejścia

Rys. 2.6. Schemat neuronu
Źródło: Opracowanie własne



$i_x \; \forall x \; \in [1, 2, 3]$: dane wejściowe
$v_x \; \forall x \; \in [1, 2, 3, 4]$: neurony w warstwie ukrytej
$o_x \; \forall x \; \in [1, 2]$: dane wyjściowe

Rys. 2.7. Schemat sieci neuronowej
Źródło: Opracowanie własne

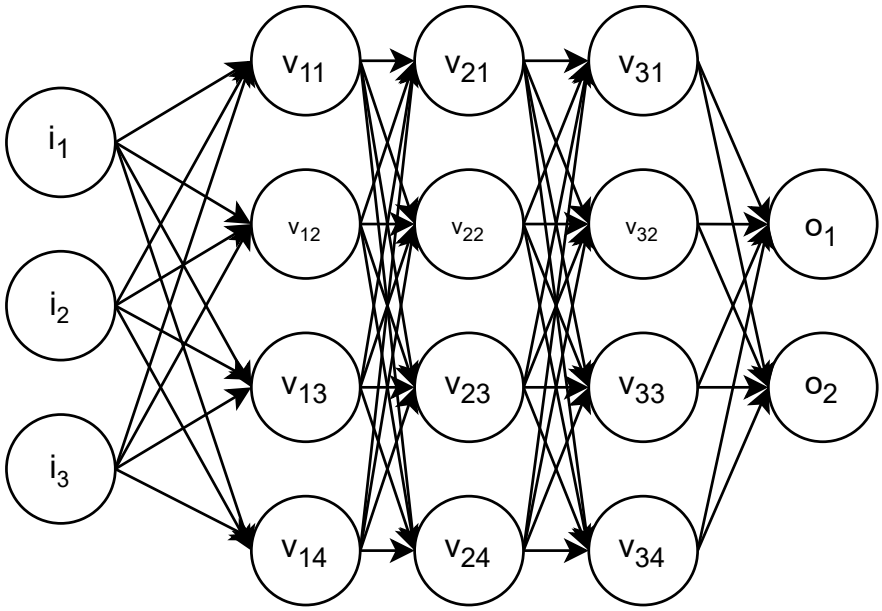
Sieci neuronowe możemy podzielić na wiele rodzajów, do których możemy zaliczyć między innymi:

- **perceptron** - jest to najstarszy przykład sieci neuronowej złożonej z jednego perceptronu (neuronu). Można je zastosować w problemach klasyfikacji;
- **sieci wielowarstwowe perceptronowe** - jest to sieć złożona z wielu warstw połączonych ze sobą neuronów, najprostszy model sieci zaprezentowany na **rysunku 2.7** . Składa się z warstwy wejściowej, warstw (jednej bądź wielu) ukrytych oraz warstwy wyjściowej. W neurony w tej sieci w porównaniu do perceptronów, mają funkcję aktywacyjną sigmoidalną, ze względu na rozwiązywanie problemów nieliniowych (posiadających więcej rozwiązań niż dwa 0/1). Można je zastosować na przykład do klasyfikacji danych;
- **sieci konwolucyjne (CNN)** - są to sieci służące do rozpoznawania obrazów, nazwa wzięła się od wykonywanej na obrazie operacji konwolucji (splotu). Sieci te posiadają dodatkowe warstwy konwolucyjne oraz spłaszczania, które pozwalają zamienić reprezentację obrazu w pojedyncz;

– **sieci rekurencyjne** - charakteryzują się pętlą zwrotną w warstwie ukrytej. Mogą być wykorzystane do generowania tekstu, tłumaczen maszynowych, a także na przykład przewidywania cen rynkowych;
– **sieci samoorganizujące się** - wykorzystuje uczenie nienadzorowane oraz. Składają sie jedynie z warstwy wejściwej i wyjściowej. Zaś cechą charakterystyczną jest to, że neurony określające podobne klasy znajdują się obok siebie. Sieci te mogą być wykrozystywane do podziału klientów na odpowiednie grupy bądź do wskazania jakim klientom zaproponować karty kredytowe [14, 15].

2.2.1. Głębokie uczenie

Jest to podkategoria uczenia maszynowego polegająca na tworzeniu wielowarstwowych sieci neuronowych. W porównaniu do podstawowych sieci neuronowych potrzebuje ogromnych zbiorów danych do utworzenia modelu predykcyjnego. Potrzebuje również dużo więcej mocy obliczeniowej przez wzgląd na ilość warstw ukrytych, których może być dużo więcej, przykładem najprosztszej sieci głębokiej jest **obraz 2.8** .



$i_x \; \forall x \; \in [1, 2, 3]$: dane wejściowe
$v_x \; \forall x \; \in [11, 12, 13, 21, 22, 23, 31, 32, 33]$: neurony w warstwie ukrytej
$o_x \; \forall x \; \in [1, 2]$: dane wyjściowe

Rys. 2.8. Schemat prostej głębokiej sieci neuronowej
Źródło: Opracowanie własne

Warstwa wyjściowa DNN może dostarczać dane róznego formatu, może to być na przykład, tekst, liczba bądź dźwięk. Posiada również bardzo dużo zastosowań w któych skład wchodzi generowanie treści, Deepfake, analiza obrazów, wskazywanie obiektów na obrazach, projektowanie leków, czatboty. Jest to udoskonalenie podstawowych sieci neuronowych. Tak więc częśc typów sieci opisanych w **sekcji 2.2** będzie odnosić się do głębokich sieci neuronowych, należy do nich CNN[**MicrosoftDepp2023**].

3. Klasyfikacja danych

4. Podejście Low-Code

5. Microsoft Azure

6. Opis doświadczenia

7. Analiza porównawcza

8. Perspektywy rozwoju

Wykaz rysunków

Wykaz tabel

Bibliografia

[1] Microsoft. „*Microsoft Machine Learning Studio (classic)*". URL: *https://studio.azureml.net/*.

[2] Charles R Harris i in. „*Array programming with NumPy*". W: *Nature* 584 (7824 wrz. 2019), s. 356–362. DOI: *9.1038/s41586-020-2649-2*.

[3] The pandas development team. „*pandas-dev/pandas: Pandas*". Lut. 2019. DOI: *9.5281/zenodo.3509134*. URL: *https://doi.org/9.5281/zenodo.3509134*.

[4] Wes McKinney. „*Data Structures for Statistical Computing in Python*". W: 2010, s. 56–61. DOI: *10.25080/Majora-92bf1922-00a*.

[5] UNB. „*CICIDS2017 | Kaggle*". URL: *https://www.kaggle.com/datasets/cicdataset/cicids2017*.

[6] Oxford University Press. „*intelligence, n., sense 1*". W: lip. 2023. DOI: *10.1093/OED/3757635879*.

[7] „*Artificial Intelligence in Science*". W: *Artificial Intelligence in Science* (czer. 2023). DOI: *10.1787/A8D820BD-EN*.

[8] Batta Mahesh. „*Machine Learning Algorithms-A Review*". W: *International Journal of Science and Research* (2018). ISSN: 2319-7064. DOI: *10.21275/ART20203995*.

[9] Eric Vyacheslav. „*The difference between Artificial Intelligence (AI), Machine Learning*". 2023. URL: *https://www.linkedin.com/feed/update/urn:li:activity:7014145513159569408/*.

[10] Crafsol Technology. „*Semi-Supervised Learning and its Application*". URL: *https://www.linkedin.com/pulse/semi-supervised-learning-its-application-crafsol-technology/*.

[11] Sun-Chong Wang. „*Artificial Neural Network*". W: *Interdisciplinary Computing in Java Programming* (2003), s. 81–100. DOI: *10.1007/978-1-4615-0377-4_5*.

[12] Robert Koch. „*History of Machine Learning - A Journey through the Timeline*". 2022. URL: *https://www.clickworker.com/customer-blog/history-of-machine-learning/*.

[13] Alexander L. Fradkov. „*Early History of Machine Learning*". W: *IFAC-PapersOnLine* 53 (2 sty. 2020), s. 1385–1390. ISSN: 2405-8963. DOI: *10.1016/J.IFACOL.2020.12.1888*.

[14] IBM. „*What are Neural Networks?*" URL: *https://www.ibm.com/topics/neural-networks*.

[15] Karolina Bartos. „*SIEĆ SOM JAKO PRZYKŁAD SIECI SAMOORGANIZUJĄCEJ SIĘ*". W: (2012). ISSN: 1507-3866.